Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Here are some of the inferences I made my analysis of categorical variables from the dataset on the dependent variable (Count)

1.Fall has the highest median, which is expected as weather conditions are most optimal to ride followed by summer.

2.Median bike rents are increasing year on as year 2019 has a higher median then 2018, it might be due the fact that bike rentals are getting popular and people are becoming more aware about environment.

3.Overall spread in the month plot is reflection of season plot as fall months have higher median.

4.People rent more on non-holidays compared to holidays, so reason might be they prefer to spend time with family and use personal vehicle instead of bike rentals.

5.Overrall median across all days is same but spread for Saturday and Wednesday is bigger may be evident that those who have plans for Saturday night not rent bikes as it a non-working day.

6.Working and non-working days have almost the same median although spread is bigger for non-working days as people might have plans and do not want to rent bikes because of that

7.Clear weather is most optimal for bike renting, as temperate is optimal, humidity is less, and temperature is less.

2. Why is it important to use drop_first=True during dummy variable creation?

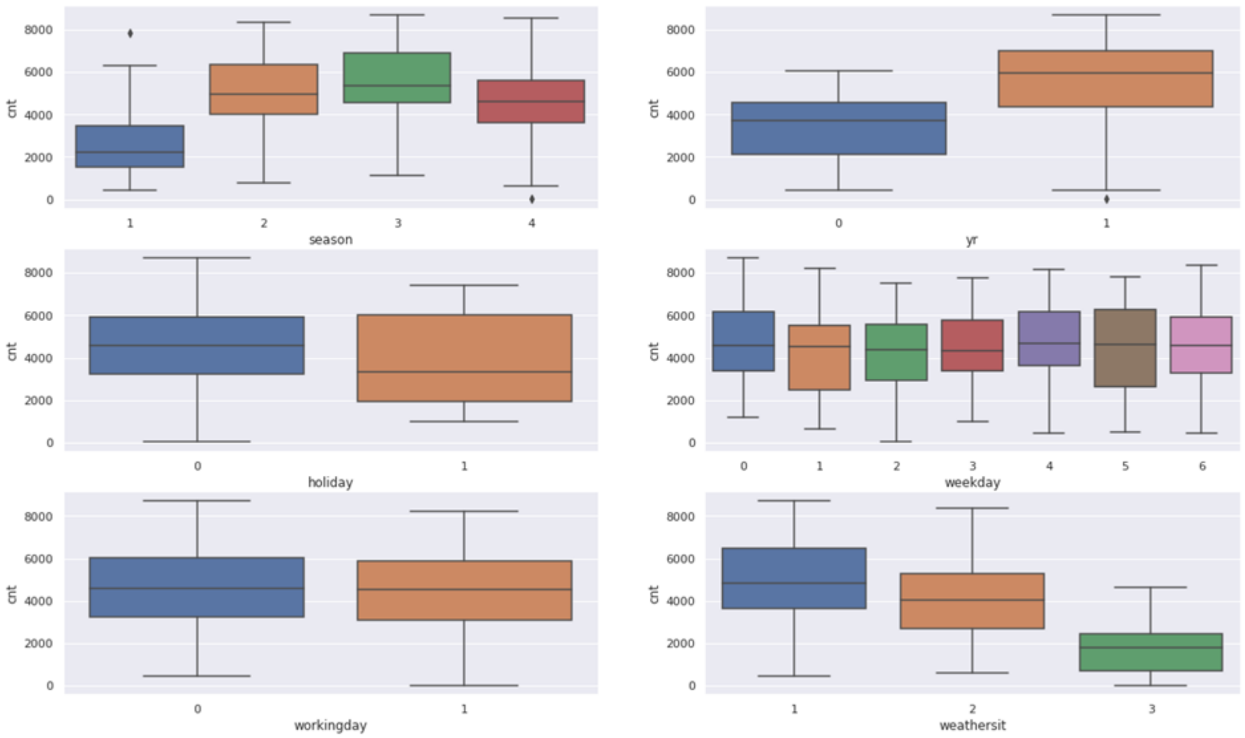
Answer. drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

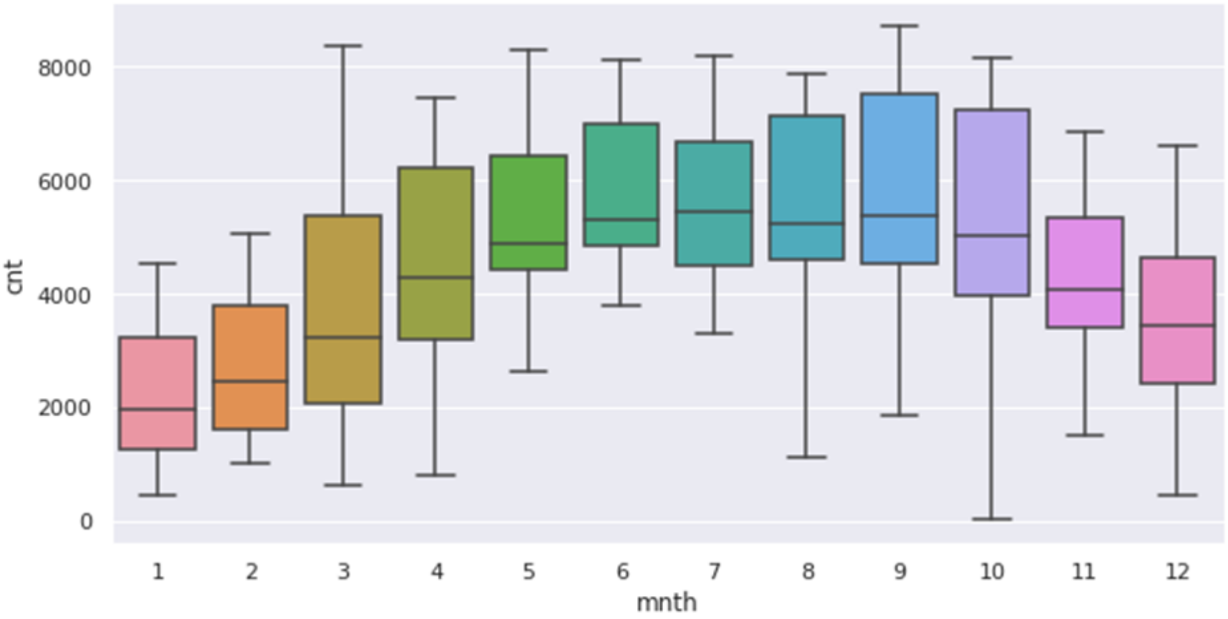
Ans.) By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

(B.) Visualising Categorical Variables:

```
plt.figure(figsize=(20, 12))

plt.subplot(3,2,1)

sns.boxplot(x = 'season', y = 'cnt', data = bb)

plt.subplot(3,2,2)

sns.boxplot(x = 'yr', y = 'cnt', data = bb)

plt.subplot(3,2,3)

sns.boxplot(x = 'holiday', y = 'cnt', data = bb)

plt.subplot(3,2,4)

sns.boxplot(x = 'weekday', y = 'cnt', data = bb)

plt.subplot(3,2,5)

sns.boxplot(x = 'workingday', y = 'cnt', data = bb)

plt.subplot(3,2,6)

sns.boxplot(x = 'weathersit', y = 'cnt', data = bb)

plt.show()
```



```
plt.figure(figsize = (10, 5))

sns.boxplot(x = 'mnth', y = 'cnt', data = bb)

plt.show()
```

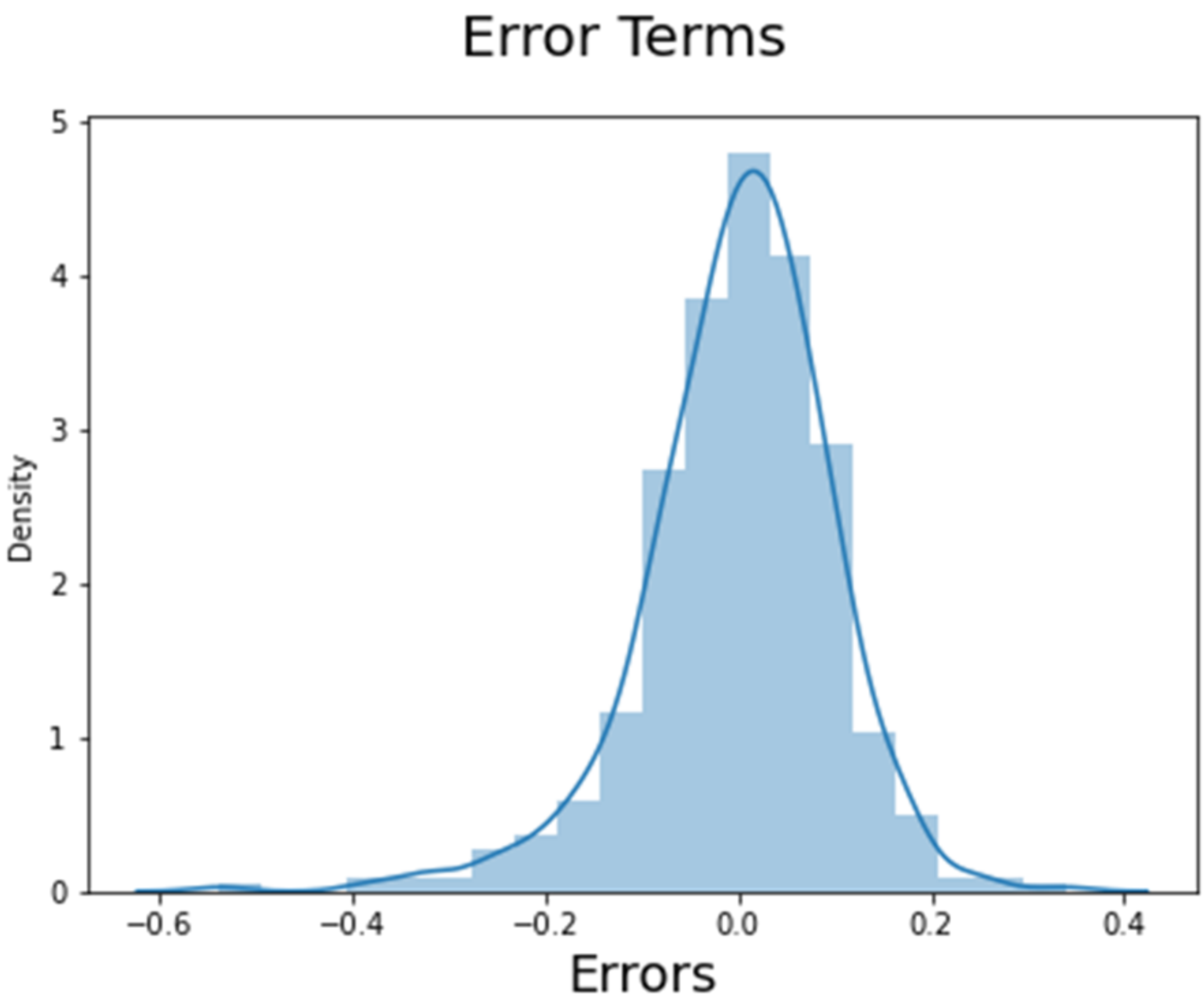


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Error terms are normally distributed with mean zero (not X, Y)

Residual Analysis Of Training Datay_train_pred = lr8.predict(X_train_lm8)

```
# residual calculations

res = y_train - y_train_pred


# Plot the histogram of the error terms


fig = plt.figure(figsize=[7,5])

sns.distplot((res), bins = 20)

fig.suptitle('Error Terms', fontsize = 20)    # Plot heading

plt.xlabel('Errors', fontsize = 18)

plt.show()
```



INSIGHT: - From the above histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer : The Top 3 features contributing significantly towards the demands of share bikes are:

weathersit_Light_Snow(negative correlation).

yr_2019(Positive correlation).

temp(Positive correlation).

General Subjective Questions with Answers

1.Explain the linear regression algorithm in detail.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

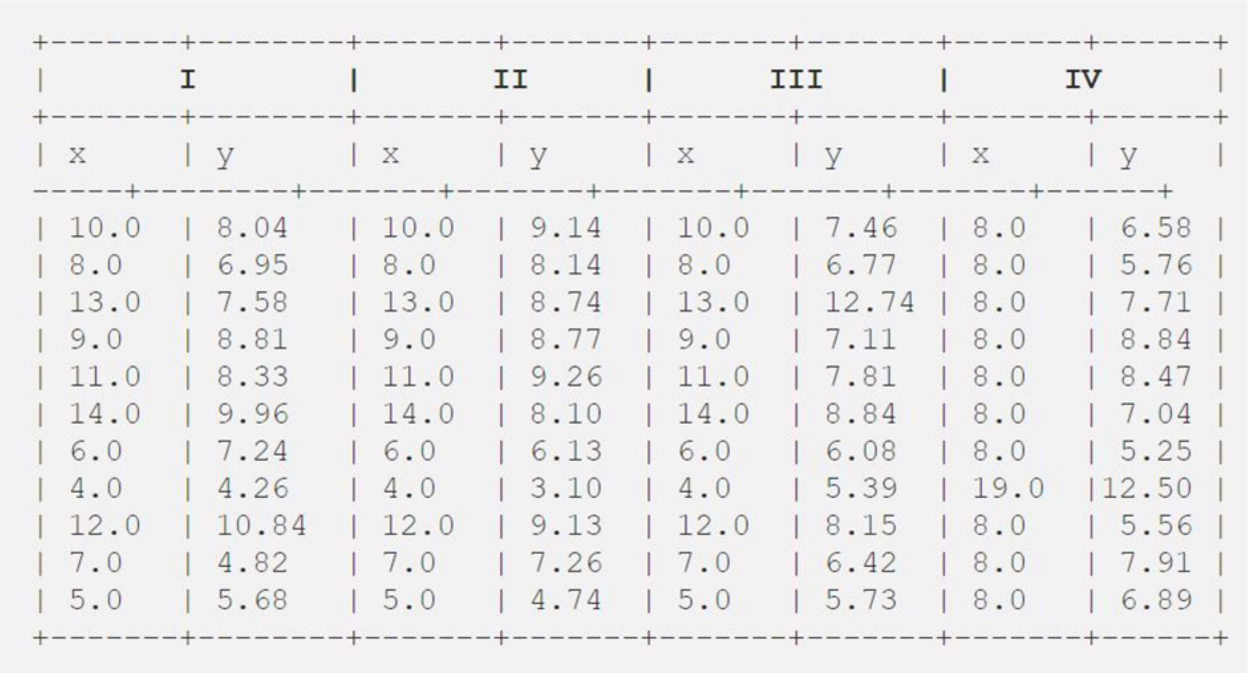
y = Dependent variable from dataset

2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Simple understanding:

Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

After that, the council analysed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

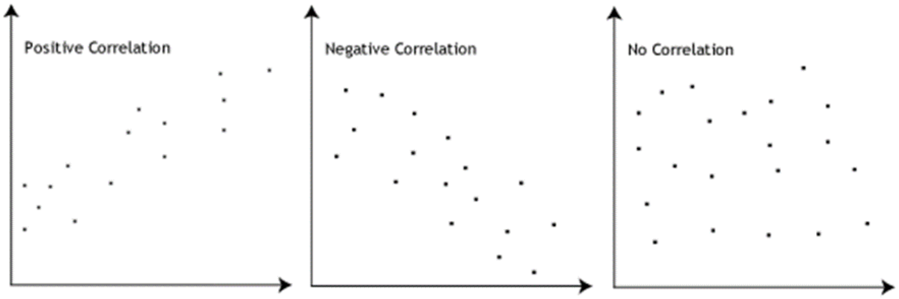
3. What is Pearson’s R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

Pearson r Formula



Pearson r Formula

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}}$$

r=correlation coefficient

x_{i}=values of the x-variable in a sample

\bar{x}=mean of the values of the x-variable

y_{i}=values of the y-variable in a sample

\bar{y}=mean of the values of the y-variable

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

	Normalisation	Standardisation
	maximum value of features are used for scaling	ndard deviation is used for scaling.
	n features are of different scales.	en we want to ensure zero mean and unit standard
	between [0, 1] or [-1, 1].	led to a certain range.
	cted by outliers.	s affected by outliers.
	provides a transformer called MinMaxScaler for .	provides a transformer called StandardScaler for n.
	mation squishes the n-dimensional data into an unit hypercube.	the data to the mean vector of original data to the ishes or expands.
	en we don’t know about the distribution	en the feature distribution is Normal or Gaussian.
	lled as Scaling Normalization	lled as Z-Score Normalization.

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen?

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

X_1=C+ α_2 X_2+α_3 X_3+⋯

〖VIF〗_1=1/(1-R_1^2)

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

X_2=C+ α_1 X_1+α_3 X_3+⋯

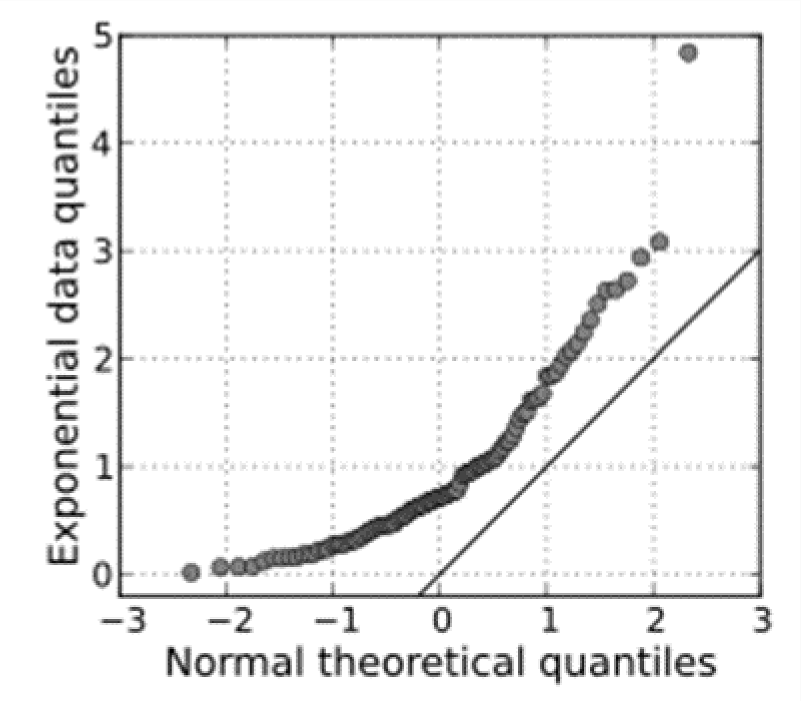
〖VIF〗_2=1/(1-R_2^2)

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.